Multi-domain Research Data Description

Fostering the participation of researchers in a ontology-based data management environment

João Aguiar Castro
Faculdade de Engenharia da Universidade do Porto / INESC TEC
Portugal
joaoaguiarcastro@gmail.com

ABSTRACT

The fast-paced production of research data in recent years has increased the awareness of the scientific community towards Research Data Management. As more and more research data is created, researchers are expected to deal with its management as soon as possible – meaning that they have to assure data preservation, sharing and reuse. Providing metadata records that meet data management requirements is as important as it is hard. Researchers are not expected to have data skills and the available tools, such as metadata standards, aren't easy to adopt, hence requiring the expertise of data professionals. This proposal addresses the problem of providing researchers with an environment where they are motivated to use the descriptors that best fit their metadata creation needs. To do so the work relies on ontologies for the representation of each research domain. This approach is expected to overcome the problems held by complex metadata schemas and favour the reuse of existing models. A panel of researchers from diverse domains is involved in the definition of application profiles and in the evaluation of the proposed methods.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.5 [**Online Information Services**]: Data sharing

Keywords

Research data management, ontology, research data description, data repository, metadata

1. INTRODUCTION

Driven by technology, research data has been produced in large scale, which represents a challenge to their management, particularly in research groups with access to limited resources [3]. The expression "data deluge" has even been adopted in the literature to express this notion [14]. Among other examples the National Science Foundation (NSF) requires that data management plans are in place to grant research projects [10], and the same happens with projects funded by the European Community [11]. Researchers are therefore driven to actively contribute to their data management activities, even before they start data production [15].

While big sciences are already supported by sophisticated tools and repositories to properly manage research data, in the context of the long tail of science [13] the problem is still to be resolved, despite the effort that has been applied by several research teams to evolve research data management.

The following section reports on research data management, and addresses the main issues surrounding metadata, highlighting existing platforms that are being implemented to tackle RDM challenges. This document also presents the thesis work to be developed – the scope is drawn, along with research questions in Section 3. Section 4 describes the research that has been done so far to support this proposal, in favour of the adoption of ontologies for research data description. Since this work depends on the cooperation in different domains, the researchers panel is also presented in this paper. This proposal is completed with an overview of the tasks and a workplan.

2. RESEARCH DATA MANAGEMENT

Although research data management it is quite a common and accepted concept in the digital repositories area, its practical application is not consolidated and it is segmented in several initiatives.

Some issues are being identified in research data management studies, such as the lack of formal data management plans, which result in researchers applying ad-hoc practices to manage their data. Additionally there is also a lack of guidelines and standardized procedures in creating and storing data, and the metadata that has been produced is minimal [19].

Research data is very heterogeneous and its content can be complex, so metadata is necessary for data interpretation. We can assume that metadata its a key aspect in research data management. Describing a dataset is essential from a data preservation perspective – if data can be interpreted by third parties, it is more likely to be reused [2]. Unlike publications, research data does not provide information about its

Digital Libraries 2014 Doctoral Consortium September, 8, 2014, London, UK

nature, which narrows the potential of its content indexing and later retrieval. However, comprehensive and accurate metadata is not easy to obtain.

2.1 Metadata for Research Data Management

To cope with metadata issues a wide variety of generic metadata schemas are currently in use, such as Dublin Core[1], and MARC[2]. A generic approach to metadata can ensure its interpretation in a wider environment and encompass the representation of a larger number of datasets, enabling interoperability [20].

Other metadata schemas are more specific and tailored for e-science applications. This is the case of the Ecological Metadata Language (EML) – a model for formalizing and standardizing the set of concepts that are essential for describing ecological data [9], and the Data Documentation Initiative (DDI) – a metadata specification for the social and behavioral sciences[3]. The inclusion of domain-specific descriptors is convenient, as they can yield higher quality metadata, however at the expenses of more time dedicated to data description tasks.

Considering the rigid nature and the growing complexity of standardized metadata schemas, a practical approach is to select a set of metadata descriptors, from diverse sources, to better fit particular applications. This "mixing and matching" approach yielded the notion of Application Profile [12]. Applications Profiles are believed to be better at capturing the life-cycle of a resource.

As data creators, researchers are able to supply better metadata records. Thus, adequate and comprehensive metadata production is strictly dependent on to their motivation and availability – the effort to describe research data is a limitation for researchers willingness to share it [1]. Furthermore it is desirable to capture metadata from an early point in the research workflow, when researchers are in the process of creating data and are more likely to have full knowledge of their production context. For data description activities, researchers typically rely on metadata standards, which are often too complex given their orientation to fulfill all the metadata goals – such as administrative tasks – thus requiring highly specialized professionals to deal with it [17]. Hence researchers can be reluctant to adopt metadata creation routinely.

2.2 Research Data Management platforms

Data repositories as integrated and robust platforms for storage, preservation and dissemination, can be seen as an opportunity for cost-effective research data management. Accordingly, various initiatives related to digital repositories tend to adopt strategies that ensure the principles of data integrity throughout its life-cycle.

Current initiatives related to data repository services are aligning metadata creation best practices with the research data life cycle. The Australian National Data Service (ANDS)[4] has adopted the ISO RIF-CS as a data interchange format and the Data Archiving and Network Services (DANS)[5] recommends Dublin Core and DDI schemas, or others, depending on the research domain. On the other hand the DataONE[6] and EDINA[7] repositories implemented research data management plans to promote data description earlier in the research workflow, and also recommend standards. Other well-documented examples include the Dryad Application Profile[8], and the institutional data repositories at the University of Edinburgh [18].

Yet, institutions often lack the infrastructure or struggle to secure data management activities, so they look forward to implement open-source, community-driven repositories [6], particularly considering that its harder to preserve, find and reuse data in the long tail of science. Therefore, the number of research data platforms has grown significantly in recent years. Among these we can highlight solutions like Figshare[9], Zenodo[10] and the CKAN[11], based on their popularity and adoption.

Figshare aims to provide an environment where authors gain scientific visibility through citation, by allowing the identification of each researcher, in a process that can be seen as agile. Zenodo is a multidisciplinary repository, based on the Invenio framework, supported by the European Organization for Nuclear Research (CERN), in cooperation with the EU FP7 project OpenAIREplus[12]. CKAN claims to be a complete out-of-the-box software solution that makes data accessible by providing tools to streamline publishing, sharing, finding and using data[13], and it has been used by several institutions[14]. CKAN started as a catalog management tool to discover open data and currently has more than 90 active contributors[15].

2.3 The Dendro Research Data Management platform

In this context a collaborative research data platform for small research groups, named Dendro, is being developed at our research group. Since data repositories usually target data at the end of research process, this platform is designed to support data description from the moment its starts to be produced, recognizing that the absence of timely description at an early stage can compromise metadata accuracy [16]. Dendro's goal is to support researchers in data management activities from the first stages of a research project, enabling an incremental data description approach. Dendro is built on ontologies and uses them as the sources for the descriptors used to meet metadata requirements from distinct research domains.

[1] http://dublincore.org/
[2] http://www.loc.gov/standards/marcxml//
[3] http://www.ddialliance.org/
[4] http://ands.org.au/index.html
[5] http://www.dans.knaw.nl/en
[6] http://www.dataone.org/
[7] http://edina.ac.uk/
[8] http://wiki.datadryad.org/Metadata_Profilehttp://edina.ac.uk/
[9] http://figshare.com/
[10] http://http://zenodo.org/
[11] http://ckan.org/
[12] http://zenodo.org/faq
[13] http://ckan.org/about/
[14] http://ckan.org/case-studies/
[15] https://www.ohloh.net/p/ckan

At the time of the deposit, is not mandatory for researchers to provide a detailed resource description, since Dendro targets an audience without data management skills in general, and for this reason, only basic descriptors are presented to the user. Later on, metadata can be refined by using richer descriptors presented as *recommendations*.

In short, Dendro can be interpreted as a staging area for later deposit in research data platforms, such as Zenodo, Figshare or CKAN [8] [7].

3. THESIS PROPOSAL

Framed in the research data management context discussed in Section 2, this work will be focused on providing insight on how to tackle some of the challenges that were identified. Particularly, the thesis research aims to provide accurate descriptors to encompass metadata needs from several research domains. We expect to foster researcher engagement in metadata creation by envolving them in the definition of their metadata models.

As data curators, our contribution with data management skills can be of great value, although the domain knowledge lies within the researchers themselves. So collaboration between data curators and researchers is paramount. Ideally data curators should accompany the research workflow, but we have to recognize that this is not feasible in the long tail of science context.

This means that researchers must be empowered for providing data description, as long as the right conditions for them to do so are in place - one must take into account that researchers can be reluctant to spend much of their time on data description activities or in continuous meetings with data professionals. In this sense a possible approach is to support researchers with a set of descriptors suggested by a data curator so they can provide metadata while producing data, postponing curator intervention until later in the workflow. At this point some research questions were defined to conduct this work.

Research Questions

- Which are the requirements that a research data management tool must ensure in order to meet researchers needs?
- Is the data repository platform proposed to deal with research data management a adequate solution to accompany researchers data management activities daily?
- In which way metadata standards and ontologies can be matched in order to provide a set of, comprehensive and accurate, descriptors that are suitable to represent research data from multiple domains?

4. ONGOING RESEARCH

Preliminary work has already been done to identify research data management needs in two research groups. Their research workflow was captured along with the definition of appropriate descriptors for metadata representation in both cases. This was achieved in a series of meetings that also enable us to run usability tests on data curation tools [6].

The first step to determine the descriptors that best fit a specific research domain was done in collaboration with a group of researchers, from the mechanical engineering domain, and consisted in the definition of an application profile to be used to register metadata from a specific kind of research experiments. These results were then used to improve the data curation tools [4]. The approach to gather information about the workflow in this research group was supported by a script adapted from the Data Curation Profile Toolkit[16], that can be enriched to fit future research needs as the work evolves to other domains. The application profile was designed to satisfy fundamental requirements for the documentation of research data [20] – It is comprehensive, by providing sufficient descriptors for the data in the given domain; it is also simple, which is important for users without data managemens skills; it promotes data interchange among the research team, enhancing data documentation as well as the identification and organization of the datasets. The same approach was followed in a research group from the analytical chemistry domain. After completing this phase, the resulting application profiles were represented as ontologies that could be imported into Dendro´s workflow.

4.1 Ontologies as metadata models

Taking into account the limitations observed in scientific metadata schemas, the research will be oriented to the creation of ontologies that can represent the domains selected for the case studies. The data model adopted in Dendro, the platform used in this work, supports via ontologies both generic description needs and those arising from the specific needs of each research community. Ontologies favour interoperability and provide conformity to the principles of Linked Open Data.

The advantages of using ontologies for research data description is that they constitute an appropriate representation of the semantics of any specific domain, and can evolve asynchronously with contributions for other user communities. Other point that favours ontologies, when compared to metadata schemas, is that they are flexible and a incremental approach, for datasets in a fast-paced, multi-domain research environment, is convenient. The experience so far has shown that lightweight ontologies are a good solution to represent the domains of our work. They can be modelled by defining few classes and avoiding the definition of many object properties, while dispensing constraints and axioms.

Lightweight ontologies are therefore valuable candidates to support the data model of a research management system. These are easily manageable by curators and easily processable by nature. The ontology-based modelling process allows Dendro to directly ingest and process ontologies as sources of descriptors for researchers to use in the annotation of their datasets [5].

4.2 The researchers panel

It is recognized that the data creation process varies from domain to domain. Therefore, in order to achieve a wider representation of the research data environment we must seek to establish as much contact with different research groups as possible. The close work with a panel of researchers has

[16] `http://http://datacurationprofiles.org/`

provided valuable collaboration in the requirement analysis for our workflows and tools. We expect this to continue with more challenging goals in description. We also expect to take this collaboration further in groups where data is already available and there is motivation for data deposit. In these cases we can act as data curators (in an experimental basis) and collaborate in data deposit in international repositories. This will allow us further study of researchers behavior with respect to deposit and the identification of possible bottlenecks in the process.

So far we have been successful to set up a close partnership with two different groups at our institution. One research group is from the mechanical engineering domain, more specifically they run double cantilever beam experiments. The other group is from the analytical chemistry domain, and their experiments consist in determining the level of sediments in a given sample.

Apart from these two research groups we have a broader panel, that kindly accepted to participate in our activities upon request. This panel includes researchers from diverse scientific backgrounds, such as biodiversity, social and behavioral sciences, as well as several engineering domains and a astronomy laboratory. The collaboration with these researchers is to be strengthened, and the objective is to grow this panel in the following stages of the thesis. Working with these researchers´ panel can be useful for achieving our goals. Their doubts, comments and suggestions can be capitalized in opportunities to improve our research data management proposals, whether the data management platform or the metadata recommendations. We hope that as this panel expands, possibly with groups from other domains, we will get further insight on the data management practices across our institution.

5. WORK PLAN

In this section we outline the plan for the thesis work. Since the methodology to be adopted evolve from the research work that has been done so far, the plan is designed as a set of iterations where work will progress incrementally.

The workplan is structured as a set of tasks to be refined as the work progresses. Some tasks have already been developed, some are still open and some tasks are refinements of preliminary experiments.

1. State of the art review and synthesis. This task is underway. A preliminary state-of-the-art report was written and is currently under revision with recent work and a more focused perspective.

2. Selection of a set of case studies with research groups in different domains. This process is running for more than a year and has resulted in close work with 2 research groups; more case studies are expected. Throughout this activity it is intended to capture an overview of the current research data management activities and data workflows in the research groups under scope. To surpass the limitations in terms of researchers availability to collaborate on a daily basis, since this is a time consuming activity, the domains can be studied by applying qualitative content analysis of researchers publications—or others types of documentation produced by a researcher in a specific context. From this, key concepts from documented experiments can be extracted to elaborate conceptual maps to support the ontologies modelling process. Interview guidelines were designed, and have been used in the interaction with researchers and will be refined, for case study evaluation, data curation and tool evaluation meetings.

3. Ontology design for the case studies. Two profiles have been defined in case studies and a third one is underway. One of the major challenges in digital repositories lies in consolidating of the descriptors to use, and to predict the difficulties in adopting common descriptors across multiple domains. It is understood that common descriptors are not able to convey all the metadata requirements for such a diversity of domains. The ontologies will be design as an outcome of the defined application profiles. At an initial phase it is crucial to develop a deeply study on ontology representation. For instance it is important to evaluate ontologies that are being proposed to deal with research data management, and also matching them with scientific metadata standards, especially the ones that targets scientific processes. Once the ontologies are designed, and validated by the researchers, and then represented as ontologies, they will be uploaded to the Dendro platform as a source of descriptors.

4. Supporting researchers in the process of data deposit. This task will be initiated in the next academic year. The main goal is to surpass the limitations that the researchers may have when they start to use the Dendro platform. This activity will also provide the necessary background for the evaluation of the application profile and the deposit process.

5. Evaluation of the application profile and the deposit process with the researchers. The evaluation has already been performed based two preliminary data management platforms [6], and will be continued using the Dendro platform.

Acknowledgements

This work is supported by project NORTE-07-0124-FEDER-000059, financed by the North Portugal Regional Operational Programme (ON.2–O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

6. REFERENCES

[1] D. Akmon, A. Zimmerman, M. Daniels, and M. Hedstrom. The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs. *Archival Science*, pages 1–22, 2011.

[2] C. L. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 2012.

[3] C. L. Borgman, J. C. Wallis, and N. Enyedy. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries.

International Journal on Digital Libraries, 7(1-2):17–30, July 2007.

[4] J. A. Castro, J. R. da Silva, and C. Ribeiro. Designing an Application Profile Using Qualified Dublin Core: A Case Study with Fracture Mechanics Datasets. *Proc. of the International Conference on Dublin Core and Metadata Applications*, pages 47–52, 2013.

[5] J. A. Castro, J. R. da Silva, and C. Ribeiro. Creating lightweight ontologies for dataset description Practical applications in a cross-domain research data management workflow. In *Digital Libraries*, pages 0–3, 2014.

[6] J. R. da Silva, J. Barbosa, M. Gouveia, C. Ribeiro, and J. Correia Lopes. UPBox and DataNotes: a collaborative data management environment for the long tail of research data. In *iPres 2013 Conference Proceedings*, 2013.

[7] J. R. da Silva, J. A. Castro, C. Ribeiro, and J. Correia Lopes. The Dendro research data management platform Applying ontologies to long-term preservation in a collaborative environment. In *iPRES 2014 Proceedings (to appear)*, 2014.

[8] J. R. da Silva, J. A. Castro, C. Ribeiro, and J. C. Lopes. Dendro: collaborative research data management built on linked open data. In *Proceedings of the 11th ESWC*, 2014.

[9] E. Fegraus and S. Andelman. Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, 86(3):158–168, 2005.

[10] N. S. Foundation. Grants.gov Application Guide - A Guide for Preparation and Submission of NSF Applications via Grants.gov. Technical report, 2011.

[11] H2020. Multi-beneficiary General Model Grant Agreement. Technical Report December, 2013.

[12] R. Heery and M. Patel. Application profiles: mixing and matching metadata schemas. *Ariadne*, (25), 2000.

[13] P. B. Heidorn. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*, 57(2):280–299, 2008.

[14] T. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[15] L. Lyon. Dealing with Data: Roles, Rights, Responsibilities and Relationships. Technical Report June, 2007.

[16] S. Macdonald and L. Martinez-Uribe. Collaboration to Data Curation: Harnessing Institutional Expertise. *New Review of Academic Librarianship*, 16(sup1):4–16, Oct. 2010.

[17] J. Qin and K. LI. How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure. In *Proceedings of the Internacional Conference on Dublin Core and Metadata Applications*, pages 25–34, 2013.

[18] R. Rice. Applying DC to Institutional Data Repositories. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, page 2008, 2009.

[19] R. Rice and J. Haywood. Research data management initiatives at University of Edinburgh. *International Journal of Digital Curation*, 6(2):232–244, 2011.

[20] C. Willis, J. Greenberg, and H. White. Analysis and Synthesis of Metadata Goals. *Journal of the American Society for Information Science and Technology*, 63(8):1505–1520, 2012.